WordCount with MapReduce and Hadoop

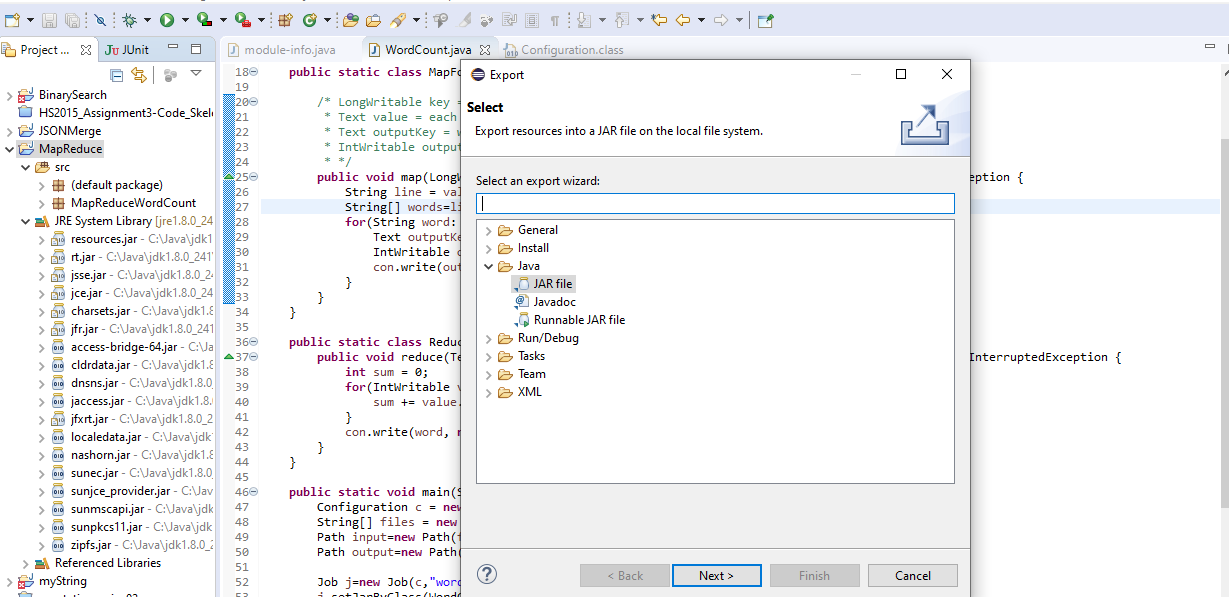
Khai Nguyen

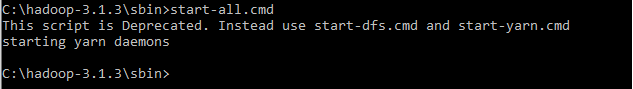
khainguyen@temple.edu

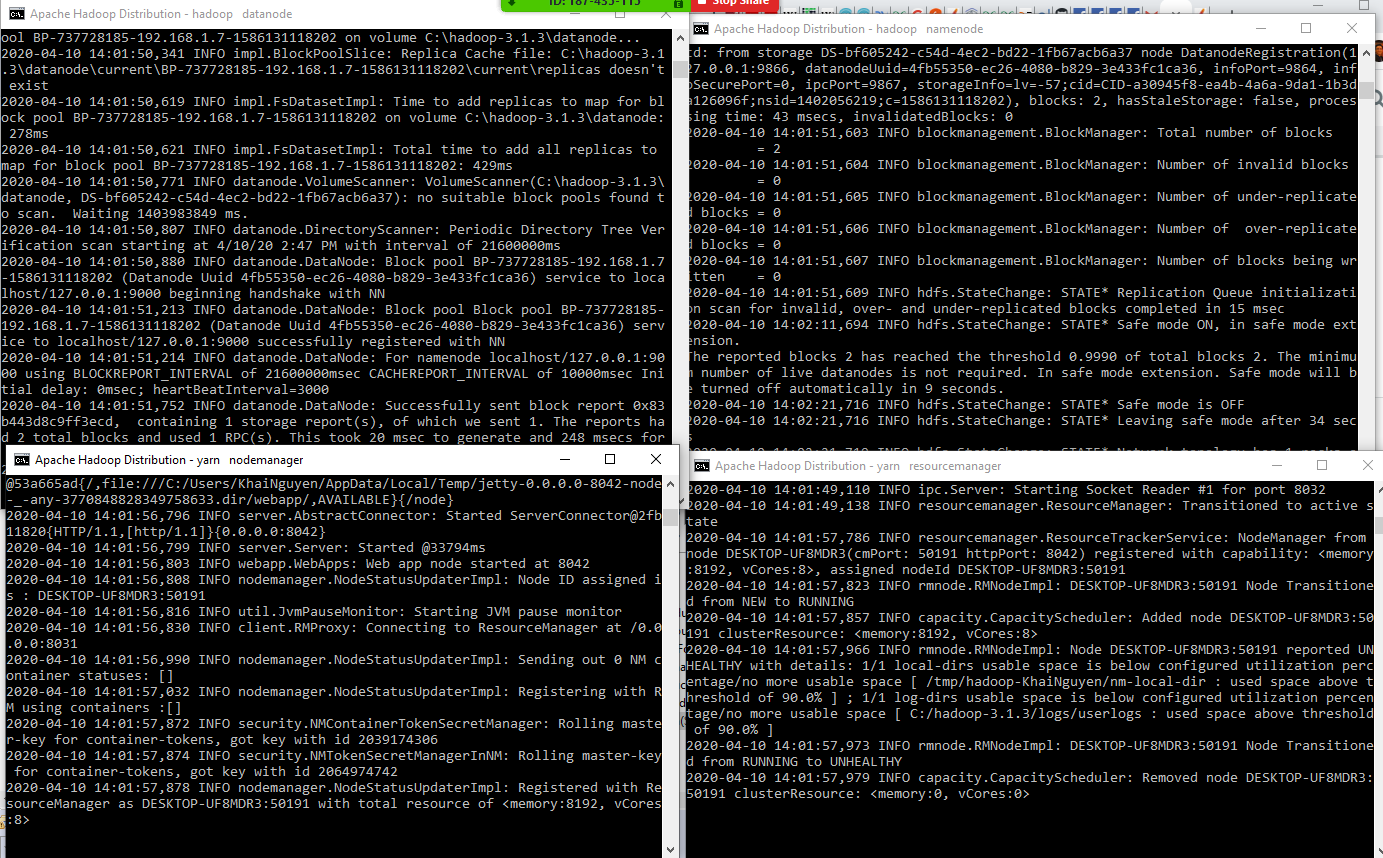
CIS 4517 Data Intensive and Cloud Computing

**Problem 1 & 2: Word Count**

To produce Project > Export > JAR File > *Next*

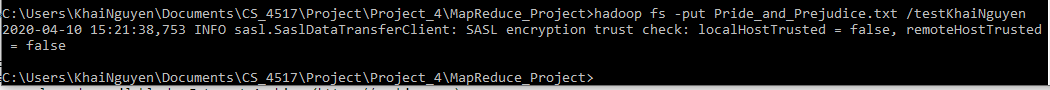






* Upload file to HDFS, store in /testKhaiNguyen

C:\Users\KhaiNguyen\Documents\CS\_4517\Project\Project\_4\MapReduce\_Project>hadoop fs -put Pride\_and\_Prejudice.txt /testKhaiNguyen



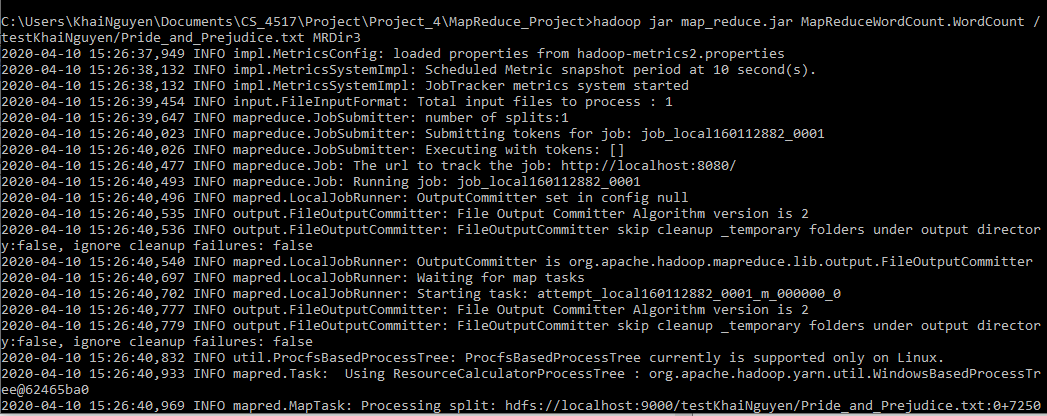
* Check if file is upload by viewing content

C:\Users\KhaiNguyen\Documents\CS\_4517\Project\Project\_4\MapReduce\_Project>hadoop fs -cat /testKhaiNguyen/Pride\_and\_Prejudice.txt

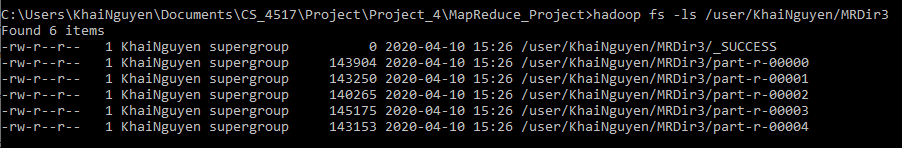


* Run MapReduce

C:\Users\KhaiNguyen\Documents\CS\_4517\Project\Project\_4\MapReduce\_Project>hadoop jar map\_reduce.jar MapReduceWordCount.WordCount /testKhaiNguyen/Pride\_and\_Prejudice.txt MRDir3



* View files created.



* Download files

hadoop fs -get /user/KhaiNguyen/MRDir3/part-r-\* .



* Sort

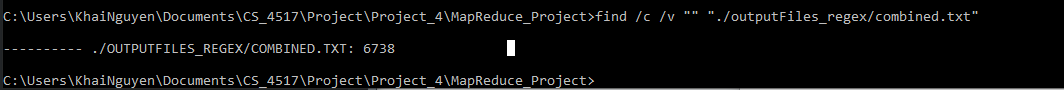
hadoop fs -cat /user/KhaiNguyen/MRDir3/part-r-\* | sort > ./outputFiles/combined.txt



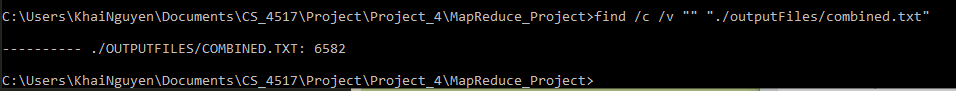
* View number of words OR number of lines since each line is designated for 1 word

find /c /v "" "./outputFiles/combined.txt"

Problem 1:



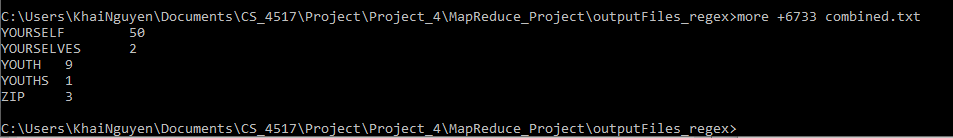
Problem 2:



* Display last 5 lines.

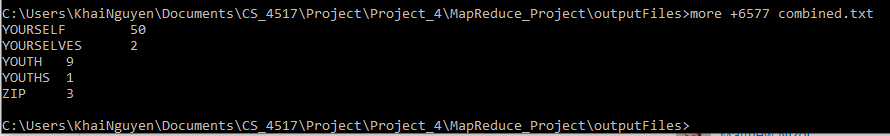
Problem 1:

Since there are 6738 lines, we can display the rest of the files starting at line 6733 to get the last 5 lines.



Problem 2:

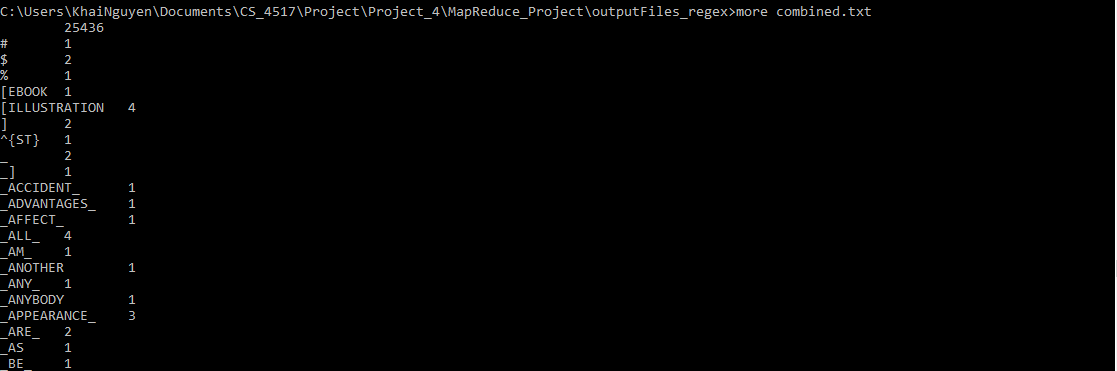
Since there are 6582 lines, we can display the rest of the files starting at line 6577 to get the last 5 lines.



* Display first line

more combined.txt

Problem 1:



Problem 2:

